Multi-Class on-Tree Peach Detection Using Improved YOLOv5s and Multi-Modal Images

LUO Qing[1,2,3], RAO Yuan [1,2,3*], JIN Xiu[1,2,3], JIANG Zhaohui[1,2,3], WANG Tan[1,2,3], WANG Fengyi[1,2,3], ZHANG Wu[1,2,3]

(1. *College of Information and Computer Science*, *Anhui Agricultural University*, *Hefei* 230036, *China*;
2. *Key Laboratory of Agricultural Sensors*, *Ministry of Agriculture and Rural Affairs*, *Hefei* 230036, *China*;
3. *Anhui Provincial Key Laboratory of Smart Agricultural Technology and Equipment*, *Hefei* 230036, *China*)

Abstract: Accurate peach detection is a prerequisite for automated agronomic management, e.g., peach mechanical harvesting. However, due to uneven illumination and ubiquitous occlusion, it is challenging to detect the peaches, especially when the peaches are bagged in orchards. To this end, an accurate multi-class peach detection method was proposed by means of improving YOLOv5s and using multi-modal visual data for mechanical harvesting in this paper. RGB-D dataset with multi-class annotations of naked and bagging peach was proposed, including 4127 multi-modal images of corresponding pixel-aligned color, depth, and infrared images acquired with consumer-level RGB-D camera. Subsequently, an improved lightweight YOLOv5s (small depth) model was put forward by introducing a direction-aware and position-sensitive attention mechanism, which could capture long-range dependencies along one spatial direction and preserve precise positional information along the other spatial direction, helping the networks accurately detect peach targets. Meanwhile, the depthwise separable convolution was employed to reduce the model computation by decomposing the convolution operation into convolution in the depth direction and convolution in the width and height directions, which helped to speed up the training and inference of the network while maintaining accuracy. The comparison experimental results demonstrated that the improved YOLOv5s using multi-modal visual data recorded the detection mAP of 98.6% and 88.9% on the naked and bagging peach with 5.05 M model parameters in complex illumination and severe occlusion environment, increasing by 5.3% and 16.5% than only using RGB images, as well as by 2.8% and 6.2% when compared to YOLOv5s. As compared with other networks in detecting bagging peaches, the improved YOLOv5s performed best in terms of mAP, which was 16.3%, 8.1% and 4.5% higher than YOLOX-Nano, PP-YOLO-Tiny, and EfficientDet-D0, respectively. In addition, the proposed improved YOLOv5s model offered better results in different degrees than other methods in detecting Fuji apple and Hayward kiwifruit, verified the effectiveness on different fruit detection tasks. Further investigation revealed the contribution of each imaging modality, as well as the proposed improvement in YOLOv5s, to favorable detection results of both naked and bagging peaches in natural orchards. Additionally, on the popular mobile hardware plat-

Received date: 2022-10-30

Foundation items: The Anhui Provincial Key Laboratory of Smart Agricultural Technology and Equipment (APKLSATE2021X004); The International Cooperation Project of Ministry of Agriculture and Rural Affairs (125A0607); The Key Research and Development Plan of Anhui Province (201904a06020056, 202104a06020012, 202204c06020022); The Natural Science Major Project for Anhui Provincial University (2022AH040125); The Natural Science Foundation of Anhui Province, China (2008085MF203)

Biography: LUO Qing (1997—), male, graduate student, research interest: smart agriculture. E-mail: tsing.omg@gmail.com

*Corresponding author: RAO Yuan (1982—), male, PhD, professor, research interest: smart agricultural information technology. E-mail: raoyuan@ahau.edu.cn

form, it was found out that the improved YOLOv5s model could implement 19 times detection per second with the considered five-channel multi-modal images, offering real-time peach detection. These promising results demonstrated the potential of the improved YOLOv5s and multi-modal visual data with multi-class annotations to achieve visual intelligence of automated fruit harvesting systems.

Key words: multi-class detection; YOLOv5s; multi-modal visual data; mechanical harvesting; deep learning

CLC number: S662.1; S126 **Documents code:** A **Article ID:** SA202210004

Citation: LUO Qing, RAO Yuan, JIN Xiu, JIANG Zhaohui, WANG Tan, WANG Fengyi, ZHANG Wu. Multi-class on-tree peach detection using improved YOLOv5s and multi-modal images[J]. Smart Agriculture, 2022, 4(4): 84-104. (in English with Chinese abstract)

罗庆, 饶元, 金秀, 江朝晖, 王坦, 王丰仪, 张武. 基于改进YOLOv5s和多模态图像的树上毛桃检测[J]. 智慧农业(中英文), 2022, 4(4): 84-104.

1 Introduction

Peach is the third most productive temperate tree species behind apple and pear, and is an excellent source of vitamin C[1]. Peach harvesting is a time-consuming and challenging task and highly dependent on labor. Efficient and easy harvesting methods are required to meet the fruit needs of a growing global population and improve orchard productivity[2]. The research and improvements of automated technology like mechanical harvesting have provided farmers with a practical approach to increase production. Traditional methods typically use segmentation algorithms or shape features such as color, shape, or texture to detect certain types of fruit[3-5]. In orchards, the detection of peaches is a challenging task due to the occlusion of branches and leaves of peach trees, as well as ever-changing illumination, which results in it being difficult to accurately detect peaches using traditional methods[6].

The first task of automated harvesting peach is to accurately detect peach. In-field fruit detection has been widely used in a variety of fruits. However, most of the images acquired in traditional methods were under controlled illumination, which makes them vulnerable to complex orchard environments[7]. Additionally, other environmental factors, such as changing appearance and morphology size of fruits, can also impose critical effect on the detection accuracy. Compared to the traditional methods, deep learning has strong adaptability to differences within a working scene, which has become one of the most promising techniques for applications in learning image features. So progressively, deep learning algorithms have been widely used in fruit detection for agricultural robots in unstructured environments[8-10]. However, in the real fruit orchards, one of the greatest challenges in fruit detection were caused by complex orchard environment, such as changing complex background[11]. Meanwhile, the varying scales of fruit targets also caused substantial difficulties in detecting fruits, especially when the fruits were bagged in orchards. Therefore, with the vigorous development of deep learning, derived from the imitation of human vision, attention mechanism was applied to enhance the model's perception ability under complex environment. In recent years, many strategies of attention mechanism have been widely adopted for various fruit detection tasks. Li et al.[12] introduced a deep learning target detection algorithm based on improved YOLOv4_tiny that combined an attention mechanism and the idea of multi-scale prediction to improve the recognition effect of occluded and small-target green pep-

pers. Jiang et al.[13] detected young apples efficiently by adding a non-local attention module and convolutional block attention model to a YOLOv4 model. Huang et al.[14] extended the target detection algorithm by adding convolutional block attention module (CBAM) to improve the performance of citrus detection. Overall, these studies demonstrated that the use of attention mechanism could enhance the model's detection performance and adapt to the natural environment with complex backgrounds.

Nevertheless, many challenging tasks still remain when seeking to effectively detect fruits in practical scenes. Under actual agricultural production environment, using RGB images as the only information to detect fruits was undesired when there were interference factors, such as occlusion or overlap of fruits and ever-changing illumination[15]. Fortunately, with the development of consumer-level RGB-D cameras, such as Microsoft Kinect and Intel RealSense, the increasing amount of information like depth data and infrared data provides us with additional cues to address these problems. Sa et al.[16] input the RGB images and infrared images to Faster R-CNN for sweet pepper identification. Fu et al.[15] developed an outdoor machine vision system with RGB-D camera to improve apple identification by using depth features to filter out the background objects. Arad et al.[17] presented a robot for harvesting sweet pepper fruits in the greenhouse. The robotic system equipped with an RGB-D camera acquired color and depth information for detecting and locating each fruit. In these aforementioned studies, it has been claimed that the introduction of more modal information in addition to RGB could contribute to the performance improvement of fruit detection. However, those studies mainly focused on detecting naked fruits without severe occlusion and overlapping because of standard planting or fruiting-wall architectures. As a matter of fact, bagging late ripening and high-quality fruits is one of the most popular ways to prevent diseases and extend storage duration[4]. This agronomic measurement increases the difficulty of in-field fruit detection because it brings more severe occlusion and irregular target shapes. Therefore, in addition to detecting naked fruits, it is also meaningful to investigate how to detect bagging peaches in an effective way.

For these reasons an efficient detection model of using three-dimensional spatial geometry and backscatter signal intensity information from multi-modal images to detect in-field naked and bagging peaches for guiding mechanical harvesting was proposed in this paper. More specifically, an RGB-D dataset of naked and bagging peaches was presented, including 4127 corresponding color, depth, and infrared images obtained by the RGB-D camera. According to the fruit picking strategy and field occlusions, the peaches were classified into four classes: un-occluded, occluded by leaves, occluded by fruits, and occluded by branches. Remarkably, the optimized detector for detecting peaches was put forward by introducing the coordinate attention mechanism and depthwise separable convolution in YOLOv5s. For purpose of evaluating the performance of the improved YOLOv5s using multi-modal images on naked and bagging peach detection and exploring the contribution of each imaging modality on environmental adaptation, abundant experiments from various aspects were implemented. Further investigation revealed the contribution of each imaging modality and the improved YOLOv5s in alleviating the negative influence of complex illumination and severe occlusion. Besides, the computational time of the proposed detection model could meet the requirements of real-time detection through its successful optimization and deployment

on NVIDIA Jetson Nano. This study might provide the possibility and foundation for performing visual intelligence in mechanical harvesting by means of utilizing the improved YOLOv5s and multi-modal visual data with multi-class annotations.

2 Materials and Methods

2.1 Data acquisition

The images acquisition was conducted using Microsoft Azure Kinect RGB-D camera (Key parameters listed in Table 1), which incorporated an RGB (Red-Green-Blue) sensor and a depth sensor that works based on the ToF (Time of Flight) principle. Data were acquired in a farming peach orchard located in Dawei Town, Hefei City, Anhui Province, China. There were two types of agronomic measurement in orchards, including naked and bagging peaches. According to the planting methods and ripening period, high-quality and late-ripening peaches were usually bagged with red papers to prevent extreme climate and disease damage. On the contrary, those early ripening peaches tended to be naked to facilitate harvesting.

Table 1 Key parameters of Azure Kinect DK camera

Feature	Parameter	Feature	Parameter
RGB camera resolution/ pix	1280×720	External dimension/ mm	126×103×39
RGB camera FOV(Field of View)/(°)	90× 59	Device interface	USB3.0
Depth camera resolution/ pix	640×576	Effective distance/m	0.25~2.88
Depth camera FOV/(°)	120×120	Ranging principle	ToF (Time of flight)

Fig. 1 shows the illustration of the data acquisition situation, on the left side are naked peach trees and on the right side are bagging peach trees.

Fig. 1 View of the naked and bagging peach visual data acquisition illustration

The RGB-D camera provides three different types of data: RGB image, IR backscattered intensity (IR), and depth image (Depth) that can be used to locate the peaches. The image data were collected in peach orchards during sunny and cloudy weather conditions. The collection periods were 7 a.m.–9 p.m. from August to September. During image acquisition, the camera was aimed perpendicular to the sunlight direction to capture the multi-modal images of peaches under normal illumination condition. The camera's viewing direction was set parallel to the sunlight direction to capture the multi-modal images of peaches under strong illumination condition. Also, the multi-modal images were gathered under artificial illumination condition during night. Considering the fact that the occlusion would affect the detection performance, some images were collected with different degrees of occluded targets from multiple viewing angles during image acquisition. According to the proportion of target area occluded by branches and leaves, the occlusion levels were classified into Slight occlusion (occluded by 0—30%), General occlusion (occluded by 30%—60%), and Severe occlusion (occluded by 60%—100%), respectively. Additionally, in order to better simulate the changing distance of the camera during mechanical harvesting, the camera was placed at the distance of 0.1 m to 1.5 m away from the tree trunk.

The distance within 0.3 m between tree trunk and the camera was considered as close distance to simulate the end-effector approaching the target. Ranging from 0.3 m to 1 m was considered as average distance to simulate the position of the camera detecting the majority of target fruits. The distance of greater than 1 m was considered as far distance to simulate the position of the camera relatively far away from target fruits.

Specific software written in C++ was developed to collect and save data automatically. The software drove the RGB-D camera to implement in-situ data recorded 5 times/s. Each time, the recorded data contained pixel-aligned one RGB image, infrared image, and depth image. In total, 4127 pairs of multi-modal images were acquired, examples were shown in Fig. 2.

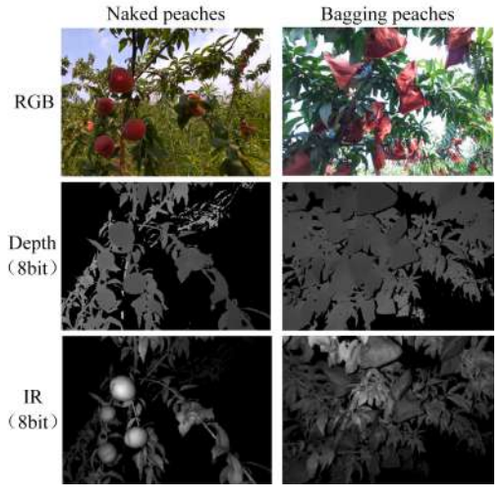


Fig. 2 Examples of multi-modal images of naked and bagging peach captured by RGB-D camera

2.2 Multi-class peach RGB-D dataset

Manual annotation was applied after the images were collected. Considering that the natural orchard existed ubiquitous occlusions among leaves, branches, and fruits. Therefore, according to the robotic picking strategy and in-field occlusion status, bounding boxes were drawn and the categories were classified into multi-class to achieve selective picking and prevent damage to the end-effectors or robots[18]. The first class indicated that the peaches were not occluded (referred to as NO in this work); the second class indicated that the peaches were only occluded by leaves (referred to as OL) and not occluded by other peaches and branches; the third class indicated that the peaches were occluded by other peaches (referred to as OF); the fourth class indicated that the peaches were occluded by branches (referred to as OB). As we know, in the process of mechanical picking, the collision between the robot arm and branches might lead to the damage of the robot arm, and the picking action of OF might cause the damage of non-target peaches. Therefore, when OB and OF appeared simultaneously for the same peach, OB was taken into considered. Additionally, and when OF and OL appears at the same time, OF was considered. For the four annotated classes, the peaches inside white, green, cyan, and brown boxes represented the NO, OL, OF, and OB, respectively.

It can be seen from Fig. 3 that all the peaches were manually labeled with bounding boxes that were tangent to peach outlines. In the case of occlusion, a peach whose occlusion area was greater than 85% and the target at the edge of the image with less than 15% area were not labelled[19]. After labeling, TXT format annotation files, including peach class names and bounding box pixel coordinates, were generated. The dataset in this research contained a total of 4127 peach images, which could be divided into two types: 2077 naked peach images and 2050 bagging peach images, respectively. This dataset has been made publicly available at https://github. com/tsing-luo/Multi-class-peach-RGB-D-dataset.

(a) Naked peaches (b) Bagging peaches

Fig. 3 Peaches were annotated into four classes, where fruits inside white, green, cyan, and brown boxes were referred to the NO, OL, OF, and OB, respectively

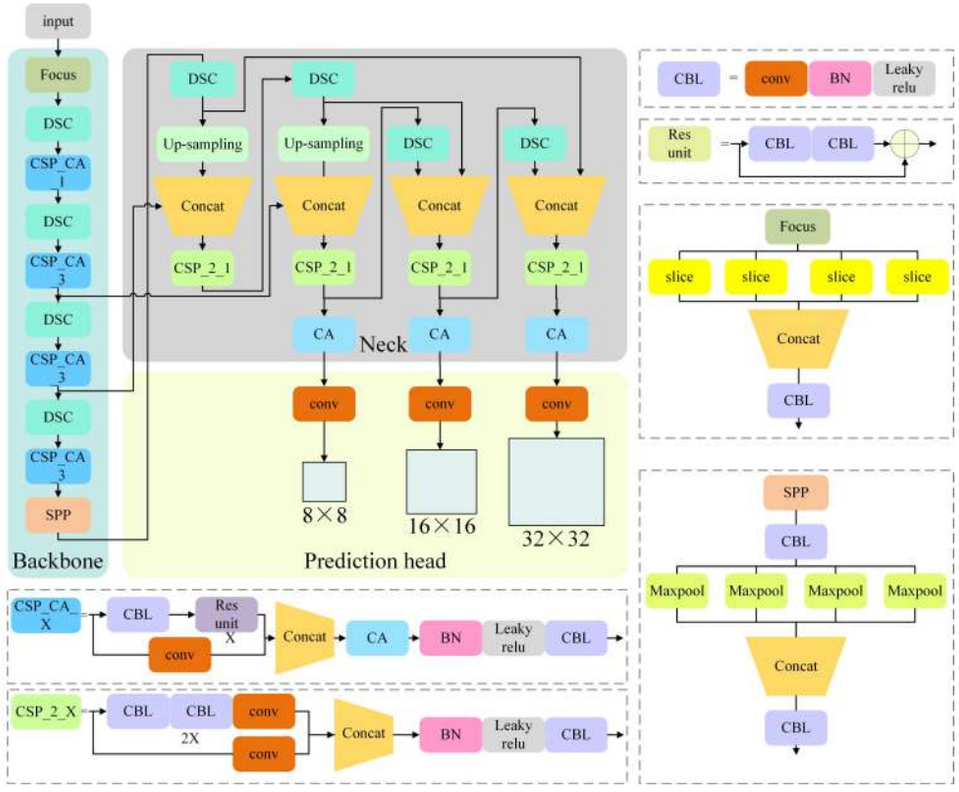
2.3 Improved YOLOv5s network

Nowadays, the YOLO[20-24] series have been one of the most popular deep learning frame works among one-stage detectors, and widely used in target detection tasks. In practical agriculture management, real-time detection under limited computation and storage resources of hardware were required, while there were limitations of the size and inference time of the fruit detection algorithms. The newly proposed YOLOv5s performed well in the pursuit of a trade-off between accuracy and speed, which could offer the fastest inference speed being up to 140 FPS (frames per second). In addition, the weight file of the YOLOv5 model was only 7.2 MB, nearly 90% less than YOLOv4. As depicted in Fig. 4, the YOLOv5s was employed as the basis of the fruit detection model in this research. The model mainly includes three parts: Backbone, Neck, and Prediction head. Its structure was modified by combining the coordinate attention mechanism and depthwise separable convolution in the backbone and neck parts.

For the original YOLOv5s model, the CSP-Darknet53 was used as the backbone network. However, due to the existence of complex backgrounds in orchards, the target features extracted from the images were easily disturbed, particularly in the case that the weeds and soil had close color to the leaves and branches of peach trees, causing the incorrect results of target detection. Meanwhile, the shallow feature map extracted from the backbone had a small receptive field that was suitable for detecting small targets, e.g., fruits[25]. Nevertheless, using low-dimensional feature maps to increase the feature information of small targets might introduce a significant amount of background noise, particularly when using multi-modal images, which might further lead to the decrease in target detection accuracy.

In order to solve these problems, as shown in Fig. 4 and Fig. 5, the CSP module design was modified in the backbone and improved the neck part by means of introducing an efficient attention mechanism, known as coordinate attention (CA), which inherited the benefits of channel attention methods while simultaneously capturing long-range dependencies with precise positional information, suppressing unimportant features and promoting useful features[26].

Previous studies have proved that adding coordinate attention to the feature extraction part of the model could enhance the representation of attention region, while adding attention mechanism to the neck part of the model could improve the position sensitivity in the detection head, preserving the relative positions between features, thus achieving



Note：CSP denotes cross stage partial layer；CA denotes coordinate attention；DSC denotes depthwise separable convolution；SPP denotes spatial pyramid pooling

Fig. 4 Overall structure of the improved YOLOv5s model

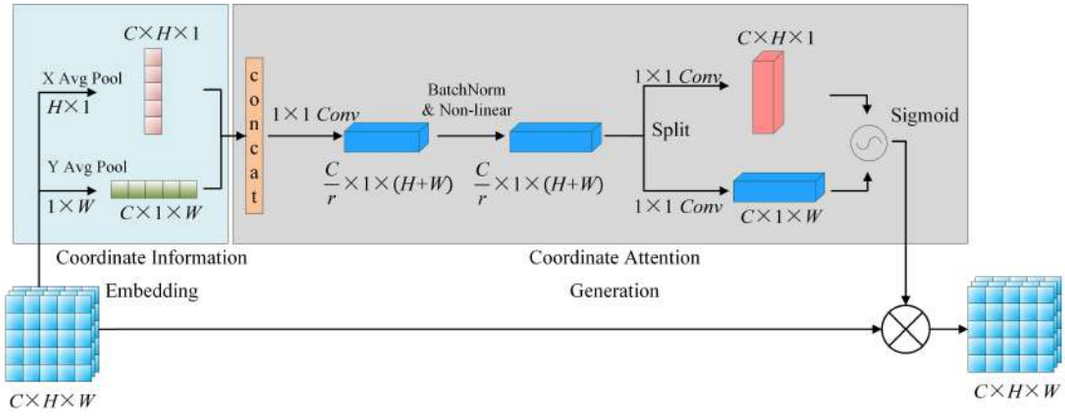


Fig. 5 Structure diagram of coordinate attention mechanism（corresponding to CA in Fig. 4）

more accurate detection results[27-29]. Specifically, given the shallow feature maps input, a pair of direction-aware feature maps were yielded by means of using two spatial extents of pooling kernels ($H\times1$) and ($1\times W$) to encode each channel along the horizontal coordinate and the vertical coordinate, respectively. These two transformations also allowed attention block to capture long-range dependencies along one spatial direction and preserve precise positional information along the other spatial direction, which helped the networks accurately locate the peaches. Then, the feature maps produced by the coordinate information embedding block were concatenated and sent to a shared 1×1 convo-

lutional transformation. The feature maps were converted to $\frac{C}{r} \times 1 \times H \times W$, r was the reduction ratio for controlling the block size as in the SE block[30]. After the feature maps underwent the Batch Norm layer and Non-linear activation function, the feature maps were split into separate tensors along the spatial dimension. Another 1×1 convolutional transformation was utilized to separately transform horizontal dimension tensors f^h and vertical dimension tensors f^w to tensors with the same channel number to the input $C \times H \times W$, the output could be formulated as Equation (1) and (2).

$$g^h = \sigma\left(C_h\left(f^h\right)\right) \tag{1}$$

$$g^w = \sigma\left(C_w\left(f^w\right)\right) \tag{2}$$

where, C denotes the convolutional transformation; σ is the sigmoid function. Finally, the output Y is written as Equation (3).

$$Y_{(i,j)} = X_{(i,j)} \times g^h(i) \times g^w(j) \tag{3}$$

Additionally, due to the limitations of hardware resources in practical agriculture management, there were requirements of optimizing the size and computational cost of the fruit detection model in addition to improving detection accuracy, which was critical for facilitating its deployment on the in-field harvesting robots. As shown in Fig. 4 and 6, depthwise separable convolution (DSC) was introduced to substitute part of regular convolutions in the backbone and neck network for reducing the model parameters and speeding up the detection inference time without penalizing the accuracy[31-33]. The DSC was a combination of depth-wise convolution and point-wise convolution. The deep-wise convolution contained c_1 convolution kernels of size $h \times w \times 1$ and achieved the filtering work by acting on each channel. The point-wise convolution contained c_2 convolution kernels of size $1 \times 1 \times c_1$ and took charge of the conversion channel by acting on the output feature map of the depth-wise convolution. Therefore, the parameters of depthwise separable convolution and traditional convolution were as follows Equation (4) and (5).

$$P_{DSC} = c_1 \times (h \times w \times 1) + c_2 \times (1 \times 1 \times c_1) \tag{4}$$

$$P_{conv} = c_2 \times (h \times w \times c_1) \tag{5}$$

By comparing the parameters of P_{DSC} and P_{conv}, it can be found out that the DSC effectively decomposed the traditional convolution by separating the spatial filtering from the feature generation mechanism.

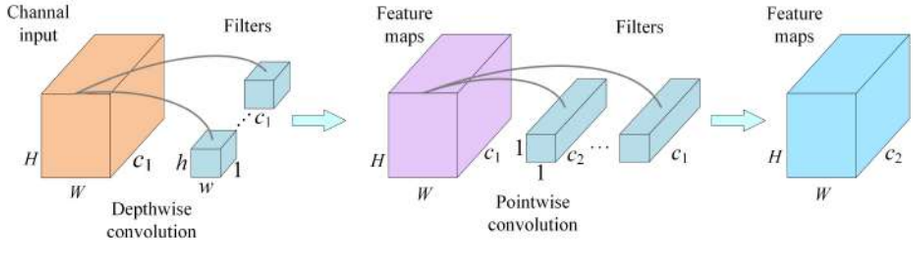


Fig. 6 Depthwise separable convolution structure (corresponding to DSC in Fig. 4)

For the specific parameters of the backbone, we chose multi-modal images with 640×640 resolution as the model input. The shallow feature information was aggregated through a Focus module, two-layer DSC, a CSP_CA_1 and a CSP_CA_3 module, the feature dimension was converted to 128×128×64. Then, additional features were extracted through the two-layer CSP_CA_3 module, two-layer DSC, and a SPP module[23]. Three adequate feature levels were obtained, the first two focused on the small-scale and medium-scale features and the last one focused on large-scale. Then, the

features were transferred to the neck part.

In the prediction head part, the k-means clustering algorithm was used to find the anchor box, and complete IoU (CIoU)[34] was used for model's bounding box regression loss, which toke three geometric properties into account, including overlap area, central point distance and aspect ratio, led to faster convergence and better performance. The formulae are as follows Equations (6)—(8).

$$CIOU_{Loss} = 1 - IOU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \quad (6)$$

$$v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \quad (7)$$

$$\alpha = \frac{v}{1 - IOU + v} \quad (8)$$

where, c represented the diagonal distance of the smallest closure area that can contained both the prediction bounding box and the ground truth bounding box; $\rho^2(b,b^{gt})$ represented the Euclidean distance between the center point of the predicted frame and the real frame; IOU represented a number from 0 to 1 that specifies the amount of overlap between the predicted and ground truth bounding box; $CIOU_{Loss}$ was used to obtain the corresponding loss.

By improving the backbone and neck network of the model and introducing CIoU loss function in the prediction head part, the size of the model was decreased, and the perception ability of the fruit detection model was improved, further enhancing the performance of detecting in-field peaches. The final output of the fruit detection model was the coordinate information of the peach targets (the prediction box of peach position) and the confidence level to a specific class, including NO, OL, OB, and OF.

2.4 Model deployment

The PyTorch framework was used to train the network and the model in PTH format was generated. After training, the model was deployed on NVIDIA Jetson Nano for further evaluating the potential of real-time detection. Jetson Nano supports TensorRT to accelerate the model, which could improve the processing speed of neural networks by optimizing the algorithm architecture. Firstly, the PTH format model was converted to ONNX format, which was an intermediate framework to bridge PyTorch model and TensorRT model. Then, the ONNX format model was converted to TensorRT format and tested on Jetson Nano. After the model was deployed on Jetson Nano, the time consumption was verified when using multi-modal images in detecting naked and bagging peaches.

3 Experimental results and analysis

To thoroughly evaluate the performance of the improved YOLOv5s using multi-modal images and explore the contribution of each imaging modality when detecting multi-class naked and bagging peaches in natural orchards, different combinations of multi-modal images were input into the improved YOLOv5s. The performance of the model was evaluated in terms of precision (P), recall (R), mean average precision (mAP), and detection speed. Firstly, for purpose of evaluating the performance of multi-modal images in the model generalization, the improved YOLOv5s was trained and validated based on different combinations of imaging modalities, and the quantitative analysis for test results on the naked and bagging peach detection was made. Secondly, to explore the contribution of each imaging modality in different orchard environments on peach detection, the detection results of naked and bagging peaches were compared and analyzed in several typical orchard scenarios, e.g., different illumination conditions, fruit occlusion levels and camera distances. Finally, the ablation study

was conducted to verify the effectiveness of the coordinate attention mechanism and the depthwise separable convolution.

3.1 Training platform and parameters

The deep learning framework used in this study was PyTorch 1.11.0. The training and testing platform included a server with an Intel Xeon Gold 5118 @ 2.30 GHz 12-core CPU, one NVIDIA RTX2080Ti (1620 MHz) GPU, with 4352 CUDA cores and 11 GB of memory running on the CentOS 7.9 system. The software tools included CUDA 11.2, CUDNN 7.6.5, and Python 3.7. Table 2 shows the network initialization parameters. All input images were adjusted to 640×640 pixels to adapt the input required for the network framework. Considering the memory constraints of the server, the batch size was set to eight in this research. 150 epochs were used to better analyze the training process. Parameters like momentum, learning rate, weight decay, and other parameters referred to the parameters in the original YOLOv5s model.

Table 2 Initialization training parameters

Input image size	Batch size	Momentum	learning rate	Decay	Epochs
640×640	8	0.9	0.001	0.0005	150

3.2 Evaluation indicators

The performance of the model was evaluated by measuring the average precision (AP), mAP, and detection speed. Among them, AP was estimated by precision (P) and recall (R), indicating the sensitivity of the network to target detection, and it was also an index that reflected the performance of the improved YOLOv5s model[35]. The P and R was defined as Equations (9) and (10).

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{9}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{10}$$

where, TP (True Positive) was the number of the targets correctly detected; FP (False Positive) was the number of the targets detected as incorrect classification; FN (False Negative) was the number of targets that were missed. The AP was defined in Equation (11), which was the area under the P and R curves. The mAP was defined in Equation (12), which was the average value of AP.

$$\mathrm{AP} = \int_0^1 P(R)\,dR \tag{11}$$

$$\mathrm{mAP} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{AP} \tag{12}$$

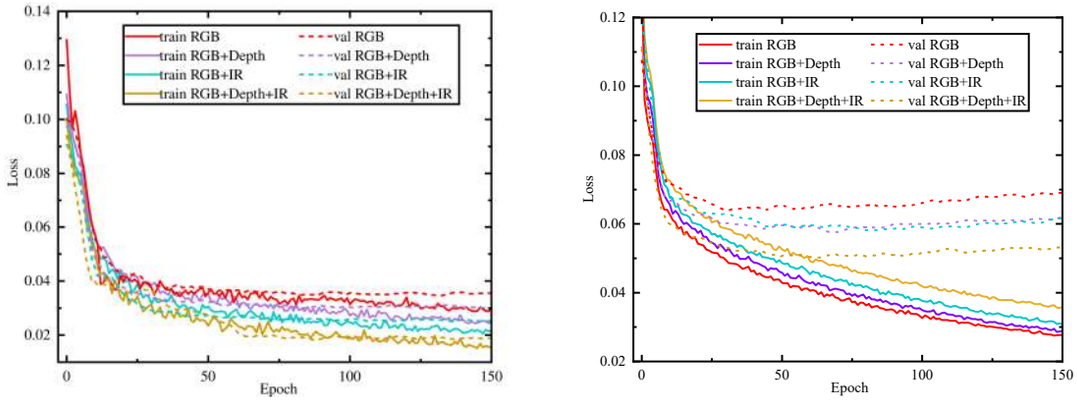
3.3 Performances of the improved YOLOv5s using different multi-modal images

In order to evaluate the performance of the improved YOLOv5s using multi-modal images in detecting multi-class naked and bagging peaches, the improved YOLOv5s was trained, validated, and tested based on different combinations of imaging modalities in this section. The dataset used in this section was the multi-class naked peach dataset and the bagging peach dataset mentioned in Section 2.2. The multi-class naked peach dataset including 2077 pairs of multi-modal images and randomly divided into three parts: training (70%), validation (10%), and testing (20%), respectively. The dataset used for training bagging peach detection model was the bagging peach dataset, with 1454, 208, and 415 pairs of multi-modal images for training, validation, and testing, respectively. The training and validation sets were applied to conduct the training of the models and determine whether and when the model started to overfit based on the training and validation curves. Then, the quantitative analysis for the test set was made to evaluate the final performance of the model using different multi-modal images. In

this study, a multi-modal image means a set of RGB, Depth, and IR images that are channel fused to obtain an image with four or five channels. For example, RGB + Depth and RGB + IR mean to stack a RGB image and corresponding Depth or IR image to obtain a multi-modal image with four channels. Similarly, the symbol of RGB + Depth + IR denotes the fusion implementation of RGB, Depth, and IR images into a five-channel image by means of channel stacking. As a result, the image number of dataset "RGB" "RGB + Depth" "RGB + IR" and "RGB + Depth + IR" were in the same, the only difference is the number of image channels in the input interface of the detection models.

3.3.1 Training assessment

From the validation curves in Fig. 7(a), it is apparent that the proposed model has not been overfitted during the training process. For training curves, the loss function reached lower values when using RGB+Depth+IR combination (plotted in brown), the fastest convergence speed appeared among all models when only using RGB modality (plotted in red). When using additional modalities like infrared (plotted in cyan) or depth (plotted in purple), the model showed lower loss values than only using RGB modality. For validation curves, it can be observed that the validation loss value and the training loss value had been very close to each other after the model converged, proving that the model learned the accurate feature information of the naked peach targets.



(a)The loss curves of naked peaches detection　　(b) The loss curves of bagging peaches detection

Fig. 7 The loss curves under different combinations of imaging modalities for naked and bagging peaches detection

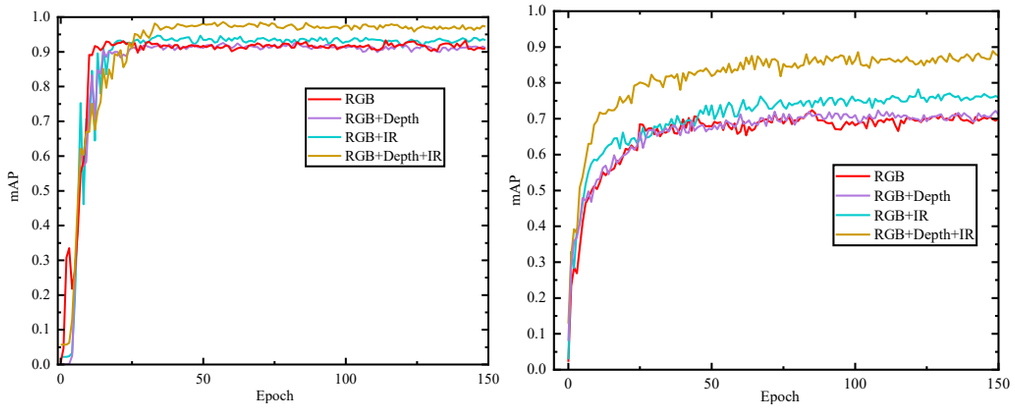
However, different from the naked peach detector, as shown in Fig. 7(b), it can be seen that the proposed model started to overfit earlier when only using RGB images than introducing some additional modalities like infrared or depth or infrared + depth. As can be seen in Fig. 7(b), the model has converged when trained to about 100 epochs, and then the model started to overfit. The fastest convergence speed appeared among all models when only using RGB modality but suffering from severe overfitting. Specifically, for training curves, the loss function reached lower values when only using RGB modality. Nevertheless, the opposite results occurred with validation losses, the reason for this phenomenon was that the overfitting of the model occurred when only using RGB modality, and the introduction of infrared and depth modalities allowed model training with stronger overfitting avoidance at the expense of a little more iterations. Compared with the naked peaches, it is apparent that the infrared and depth modalities made more contributions to improving the ability of model

generalization and overfitting alleviation, and the best result was achieved when using all imaging modalities.

Based on the above results, it can be concluded that the infrared and depth modalities really helped in improving the ability of model generalization, as expected, the best results could be achieved when using five-channel images, namely all imaging modalities.

3.3.2 Quantitative analysis for test results of different modal images

Regarding the test set, Fig. 8(a) and (b) presents the mAP of naked and bagging peaches with different combinations of imaging modalities versus epochs. Table 3 presents the naked and bagging peach detection results when using four combinations of different imaging modalities in detail. Comparing the results from RGB images and 4-channel images (Table 3, rows 1—3), the RGB images with additional infrared modality offered the best performance with the mAP of 94.7% and 78.2% for naked and bagging peaches, followed by the mere RGB images with a mAP of 93.3% and 72.4%, respectively. The least valuable combination was the RGB images with the addition of depth modality, which was even less effective than only using RGB images. Best fruit detection results were obtained when combining all modalities together, achieving the mAP of 98.6% and 88.9% for naked and bagging peaches. The most important benefit of introducing infrared and depth modalities was found out to be the precision metric in bagging peaches detection, increasing by 19.1% from 69.2% (RGB) to 88.3% (RGB+Depth+IR). This is because the extra geometric information provided by infrared and depth modalities was advantageous in reducing false positives. The recall metric also increased when introducing infrared and depth modalities, but not as significantly as the precision metric. When using the combination of RGB+Depth+IR modalities, the mAP achieved an improvement of 5.3% and 16.5% as compared to only using RGB images in detecting naked and bagging peaches.



(a) The mAP curves of naked peaches in the test set (b) The mAP curves of bagging peaches in the test set

Fig. 8 The mAP curves under different combinations of imaging modalities for naked and bagging peaches detection

Between infrared and depth modalities, it should be emphasized that the introduction of the former brought more improvement of mAP. Additionally, regarding the inference speed of the multimodal target detection model, the inference time per image only slightly increased with the increment in the number of channels. This can be explained by the fact that the increase of image channels only affected the first layer of the convolutional network, consequently the increased computation cost was

Table 3 Detection results from the test set using four combinations of imaging modalities for naked and bagging peaches

Channels	Naked peaches			Bagging peaches			Detection speed/FPS
	Precision/%	Recall/%	mAP/%	Precision/%	Recall/%	mAP/%	
RGB	90.4	92.2	93.3	69.2	71.5	72.4	**70.9**
RGB+Depth	92.1	91.5	92.7	68.7	71.2	72.3	68.0
RGB+IR	92.7	94.3	94.7	77.8	71.7	78.2	68.0
RGB+Depth+IR	**97.4**	**98.2**	**98.6**	**88.3**	**85.4**	**88.9**	66.2

negligible for the whole network. Taken together, these results demonstrated that three-dimensional spatial geometry and backscattered signal intensity information provided by infrared and depth modalities could effectively improve the fruit detection accuracy, especially in the case of implementing bagging peach detection. Last but not least, the aforementioned detection results impressed us that it was always more challenging to implement the detection of bagging peaches compared to naked peaches in orchards.

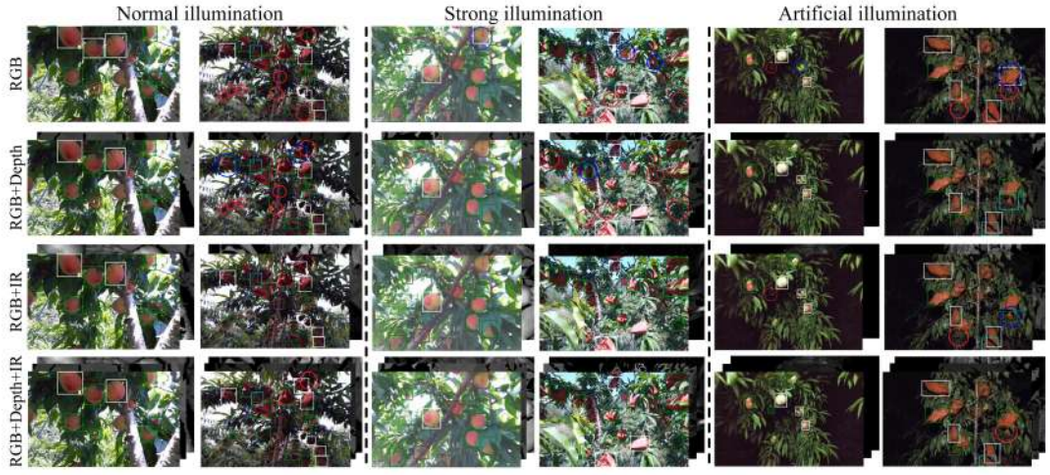
The improved YOLOv5s model was optimized by TensorRT to increase the inference speed on the Jetson Nano board. The model supports three kinds of precisions for optimization: floating-point 32 (FP32), floating-point 16 (FP16) and integer 8 (INT8). Since Jetson Nano does not support INT8 optimization, the model was converted to floating point 32 (FP32) and floating point 16 (FP16) operations, resulting in the detection speeds of 14 and 19 FPS in the test set, respectively. That's to say, the implementation of 14 and 19 times detection per second on five-channel multi-modal images. Therefore, the improved YOLOv5s model optimized by TensorRT-FP16 precision was selected for deployment in the Jetson Nano development board, which was adequate for computer-vision based peach detection and harvesting.

3.4 Contribution of different imaging modalities in typical scenarios

In order to explore the contribution of each imaging modality on peach detection under different orchard environments, the test set visualized results of naked and bagging peach detection under typical orchard scenarios were analyzed. More specifically, in the real orchard environments, there were different illumination, different occlusion levels, and different camera distance. The contribution of each imaging modality was analyzed by different combinations of imaging modalities (RGB, RGB+Depth, RGB+IR and RGB+Depth+IR) for the detection of naked and bagging peaches under different scenarios. Note that, in case of concurrent fruit occlusion, the model output will follow the labeling rules in the Section 2.2, which means that the priority sequence of fruits was OB, OF, OL and NO from highest to lowest level, respectively.

3.4.1 Comparison with different illumination conditions

Fig. 9 shows the detection results of multi-class naked and bagging peaches using different modality combinations (RGB, RGB+Depth, RGB+IR and RGB+Depth+IR) under three typical illumination conditions in the test set. The peach trees under the three conditions were at the distance of about 1 m from the camera. For each condition, four different fruit detection results were separately presented depending on the input data type: RGB (first row), RGB+Depth (second row), RGB+IR (third row), and using the modalities of RGB, Depth, and IR simultaneously (fourth row), in odd columns were naked peaches, while in even columns were bagging peaches.



Note: Missing peaches were marked in red and false detections were in blue; fruits inside white, green, cyan, and brown boxes referred to the NO, OL, OF, and OB, respectively

Fig. 9 Examples of multi-class naked and bagging peach detection results in the test set when using different modality combinations under three typical illumination conditions of Normal illumination, Strong illumination and Artificial illumination

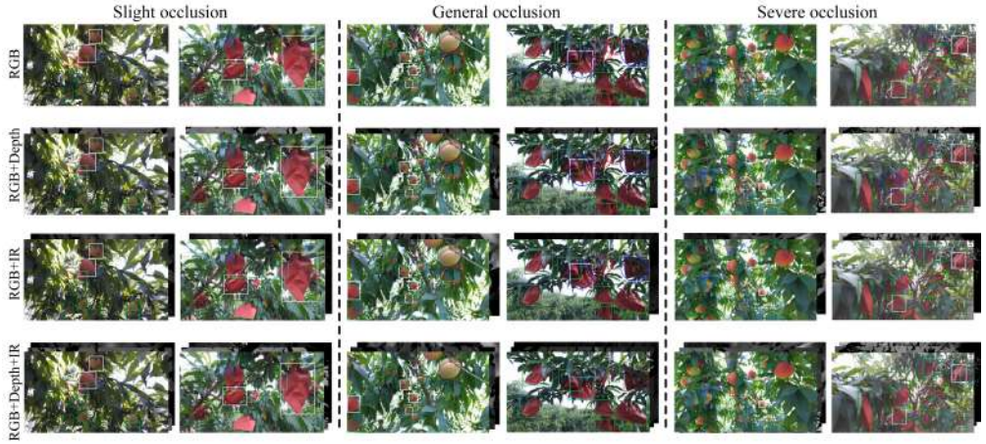
Under Normal and Strong illumination, the model detection performance when using the combination of RGB+Depth images was even worse than only using RGB images, suffering from more missing detections. The reason was the fact that ToF-based depth camera in outdoor environment was prone to noise interference due to the sunlight exposure, and the camera accuracy decreased as the measurement distance increased. As a result, there might be high possibilities that the fusion with RGB images made the detection model misjudged. Although the depth images were not suitable for peach detection in direct sunlight exposure, they did contribute to the better detection results in the artificial illumination. The explanation for this was in the nighttime, the depth camera was not interfered by sunlight noise and helped in accurately reconstructing the peach shape. Especially, when detecting some of OB and OF peaches, the RGB images presented a non-colored and invisible region of the peach edge, as well as similar color of peaches and leaves, whereas the depth images showed high distinctive geometric features. Hence, the geometric features of the OB and OF peaches appeared in the depth images were more distinguishable than those in the RGB images, making them more conductive to be successfully detected in the nighttime. Meanwhile, when comparing results before and after using infrared modality in the natural environment of daytime, a reduction in false positives of NO and OL peaches was witnessed, especially in detecting bagging peaches. One possible explanation for this might be that there were significant differences among infrared intensity of fruits and leaves in the daytime. Thus, the infrared images could effectively help in distinguishing fruits from the background in the acquired images under bright illumination condition.

Therefore, it can be concluded that the addition of depth images could help in reducing false and missing peaches under artificial illumination condition, whereas the addition of infrared images could improve peach detection accuracy in bright illumination environment compared to only using RGB images. Nevertheless, when all imaging modalities were used simultaneously, the best results could be

obtained in any illumination environments.

3.4.2 Comparison with different occlusion levels

As shown in Fig. 10, further experiments were conducted for analyzing the contribution of the different imaging modalities to multi-class naked and bagging peach detection in different occlusion levels. As mentioned in Section 2.1, based on the image's proportion of peaches occluded by branches and leaves, the occlusion levels were considered as Slight occlusion, General occlusion, and Severe occlusion.



Note: Missing bagging peaches were marked in red and false detections were in blue, fruits inside white, green, cyan, and brown boxes referred to the NO, OL, OF, and OB, respectively

Fig. 10 Examples of multi-class naked and bagging peach detection results in the test set when using different modality combinations in different occlusion scenes

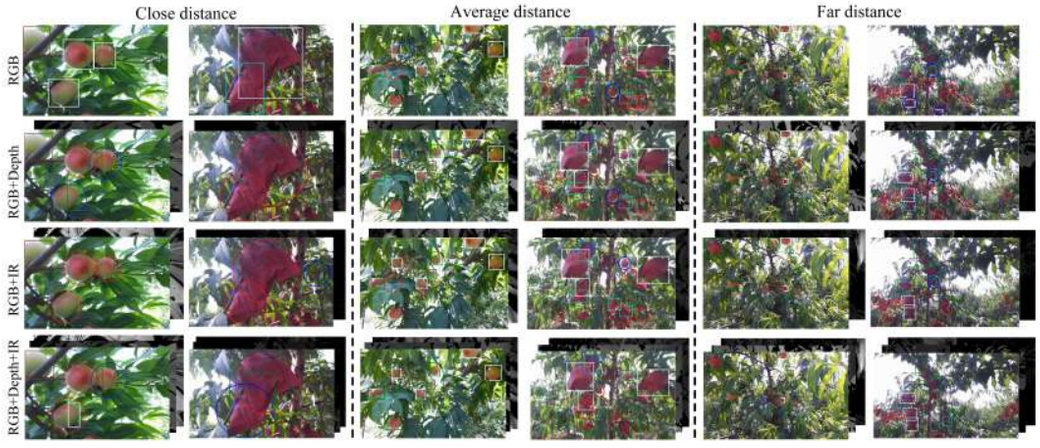
Under Slight occlusion condition (first and second columns), the peach detection missed some peaches when only using RGB images, whereas all the naked and bagging peaches were accurately detected by other detectors after further fusing Depth or Infrared images. As the bagging peaches were wrapped in similarly colored bags, such as OF peaches, several overlapping peaches were not correctly detected by the RGB detector. Meanwhile, as can be seen in the first column, naked peaches that were occluded and overlapping could be correctly detected after further fusing Depth or Infrared images. Under General occlusion (third and fourth column) status, the fusion of infrared and depth channels provided more deep features of the peach targets, and the detection results of fusing five-channel images had a more significant improvement in terms of accuracy and recall compared to those with fusing other three modalities. The depth and infrared images could offer additional fruit geometry features that differed from RGB images, such as the information of fruit edges, shape, and the distance, enabling accurate fruit detection despite of the existence of leaf or branch occlusion. It should be noted that the multi-modal images were more effective in improving the bagging peach detection accuracy than naked peaches, which could significantly reduce the rate of missing and false detections. Similarly, in the case of Severe occlusion (fifth and sixth column), although there were still cases of missing detections in detecting naked and bagging peaches even using five-channel images, the detection results were still significantly better than those using other multi-modal combinations.

Hence, it can be concluded that the introduction of infrared and depth modalities could provide

the model with more valuable information, e.g., geometric features, consequently improving the accuracy and recall rate of fruit detection, even in the case of severely occluded peaches.

3.4.3 Comparison with different camera distances

Fig. 11 presents the effect of different imaging modalities on multi-class naked and bagging peach detection with different camera distances. As mentioned in Section 2.1, the distance within 0.3 m away from the tree trunk to the camera was considered as Close distance. The distance between 0.3 m and 1 m was considered as Average distance, and the Far distance referred to the distance of greater than 1 m away from the tree trunk.



Note: Missing bagging peaches were marked in red and false detections were in blue, fruits inside white, green, cyan, and brown boxes referred to the NO, OL, OF, and OB, respectively

Fig. 11 Examples of multi-class naked and bagging peach detection results in the test set when using different modality combinations in different camera distances

The scene of the Close distance was shown in the first and second columns of Fig. 11, where some peaches were less than 0.2 m away from the camera, and the others were within 0.3 m. As can be seen, the model using only RGB images achieved the best detection results both in naked and bagging peaches, in contrast, the model received large amount of missing and false detections after fusing depth and infrared images. What's worse, the peaches within the distance of 0.2 m failed to be detected accurately. The reason was that the depth information was obtained based on the ToF mechanism in our work, however, there were operation distance requirements when using Azure Kinect DK camera. As can be seen in the first and second columns of Fig. 11, while the distance between camera and peaches was not in the camera operation distance, the depth and infrared information of the peaches will be lost, which further imposed negative effect the detection after fusion with the RGB images. Similar results occurred in Far distance detection, where some peaches far from the camera and severely occluded failed to be detected even when five-channel images were employed simultaneously. When the camera operated within the distance from 0.3 m to 1 m from the tree trunk, that's to say, at Average distances, there were best detection results when using the combination of five channels.

Therefore, there are requirements of appropriate camera operating distance if one intends to improve the detection accuracy of peaches by introducing infrared and depth modalities. In conclusion, the

introduction of infrared and depth channels definitely could improve the detection accuracy of occluded fruits, but only when the camera operated within an appropriate distance.

3.5 Ablation experiments of the improved YOLOv5s

To demonstrate the effectiveness of the improvement in YOLOv5s, this section conducts an ablation study of the improved YOLOv5s using all imaging modalities for multi-class peach detection. Specifically, the comparison contains a baseline and other three cases. The baseline model was the original YOLOv5s without attention mechanism and the depthwise separable convolution. Then, the coordinate attention mechanism and the depthwise separable convolution were integrated into YOLOv5s separately for enhancing the learning of important information, as well as reducing the number of model parameters. The network that fused DSC in the YOLOv5s were denoted as YOLOv5s-DSC, while the network that only used CA were denoted as YOLOv5s-CA. The results were compared with those models using all imaging modalities, which meant that the same RGB images training dataset as in Section 3.3, as well as their corresponding infrared and depth images, were considered.

As summarized in Table 4, four comparison experiments were carried out to investigate the performance of the CA and the DSC modules.

Table 4 Detection results of different models in the test set of naked and bagging peaches

Models	Naked peaches			Bagging peaches			Parameters/M	Detection speed /FPS
	Precision/%	Recall/%	mAP/%	Precision/%	Recall/%	mAP/%		
YOLOv5s	96.5	94.2	95.8	81.7	89.6	82.7	7.07	66.2
YOLOv5s-DSC	95.4	93.5	95.0	80.0	78.0	80.0	**5.03**	**94.3**
YOLOv5s-CA	**97.7**	**98.3**	**98.8**	**88.5**	85.1	**89.4**	7.12	66.2
Improved YOLOv5s	97.4	98.2	98.6	88.3	**85.4**	88.9	5.08	77.5

It can be seen from Table 4, when embedding the attention mechanism into the YOLOv5s, the mAP of YOLOv5s-CA was 98.8% for naked peaches, increasing by 3% as compared to YOLOv5s, outperforming all of the rest models. Unexpectedly, after substituting the regular convolution to DSC, the mAP of YOLOv5s-DSC was 95.0%, which was slightly lower than YOLOv5s. What should not be ignored was that the mAP of YOLOv5s-CA and YOLOv5s-DSC for bagging peaches was 89.4% and 80.0%, respectively. With regard to the model parameters, the YOLOv5s-DSC was 5.03 M, which was the least one among all models and 39.9% less than the original YOLOv5s. In terms of detection speed, the YOLOv5s-DSC model was 30.5% faster than that YOLOv5s, which indicated that the DSC module was more cost-effective than regular convolution. Surprisingly, the YOLOv5s-CA model's detection speed was the same as the original YOLOv5s, verifying that the coordinate attention module could enhance the feature extraction ability without significantly increasing the model parameters. After further fusing the DSC and CA module, the mAP of the improved YOLOv5s model recorded better results than the original YOLOv5s and YOLOv5s-DSC, increasing by 2.8% and 6.2% than the YOLOv5s on the naked and bagging peach detection. Note that the mAP of the improved YOLOv5s model decreased very slightly in naked and bagging peaches detection when compared to the YOLOv5s-CA. However, the improved YOLOv5s achieved 77.5 FPS detection speed with fewer parameters, which was 14.6% faster than the YOLOv5s-CA.

Overall, these experimental results demonstrat-

ed that the introduced CA and DSC was effective in improving detection accuracy and reducing computational cost of the YOLOv5s, as the proposed model could detect naked and bagging peaches in orchards with faster speed and higher accuracy while requiring fewer parameters.

3.6 Comparison and discussion

3.6.1 Comparison with other object detection networks

To further analyze the performance of the improved YOLOv5s network, the performance of the model was compared with that of three lightweight object detection networks: YOLOX-Nano[36], PP-YOLO-Tiny[37] and EfficientDet-D0[38]. The same training, validation, test sets and image channels (five-channel images, RGB+Depth+IR) were used to train and test the three networks. The detection results of the three methods on the test set are shown in Table 5.

Table 5 Detection results of different lightweight object detection models in the test set of naked and bagging peaches

Models	Naked peaches			Bagging peaches			Parameters/M	Detection speed/FPS
	Precision/%	Recall/%	mAP/%	Precision/%	Recall/%	mAP/%		
YOLOX-Nano	76.3	74.1	75.7	71.4	79.2	72.6	**0.91**	**170.2**
PP-YOLO-Tiny	78.2	79.9	80.5	80.4	78.8	80.8	1.3	154.5
EfficientDet-D0	85.3	86.9	87.7	86.2	85.1	85.4	3.9	110.7
Improved YOLOv5s	**97.4**	**98.2**	**98.6**	**88.3**	**85.4**	**88.9**	5.08	77.5

As can be seen from Table 5, improved YOLOv5s achieved best results in terms of precision, recall, and mAP compared with other three networks. The mAP of improved YOLOv5s was 98.6%, which was 22.9%, 18.1 and 10.9% higher than those of the YOLOX-Nano, PP-YOLO-Tiny, and EfficientDet-D0 in detecting naked peaches, respectively. Meanwhile, compared with other three networks in detecting bagging peaches, the improved YOLOv5s also was the best in terms of mAP, which was 16.3%, 8.1% and 4.5% higher than YOLOX-Nano, PP-YOLO-Tiny, and EfficientDet-D0, respectively. Although the average detection speed of improved YOLOv5s on the test set was 77.5 FPS, which was slower than that of other three networks, the detection accuracy was effectively improved.

From the comparison results, it can be concluded that the peach detection network based on the improved YOLOv5s proposed in this study can detect peaches more effectively and accurately than other lightweight networks.

3.6.2 Comparison with other fruit detection studies

In addition, two open-source fruit datasets including apple and kiwifruit were also performed to assess the effective of the improved YOLOv5s on the fruit detection. Specifically, Gené-Mola et al. [2] presented a Fuji apple dataset, which were acquired at night using a depth camera, and used Faster R-CNN for apple detection. Suo et al.[18] classified the kiwifruit dataset into five classes based on occlusion status and used YOLOv4 for fruit detection. Since there were different label classifications and image resolutions in these two datasets, the same label classifications and image resolutions as the raw dataset for ensuring that the comparisons were made under the same experimental conditions. Parameters like momentum, learning rate, weight decay, and other parameters referred to the parameters

in the original YOLOv5s model. Meanwhile, both the aforementioned datasets were split into training and test set, conducting the training and test of the improved YOLOv5s. Experimental results presented in Table 6 revealed that the proposed improved YOLOv5s model offered better results in different degrees than other methods in detecting fruits and verified the effectiveness on different detection tasks.

Table 6 Detection results of different models on two open-source fruit datasets.

Model	Precision/%	Recall/%	mAP/%
Faster R-CNN（Fuji apple）	84.7	88.8	92.7
Improved YOLOv5s（Fuji apple）	95.9(+8.2)	96.3(+2.5)	98.2(+5.5)
YOLOv4（Hayward kiwifruit）	——	——	91.9
Improved YOLOv5s（Hayward kiwifruit）	95.32	94.63	94.2（+2.3）

4 Conclusions

It is crucial to develop good methods of effectively detecting the fruits with different agronomic measurements for improving the popularity of mechanical harvesting. In this paper, a multi-class RGB-D dataset of natural naked and bagging peaches has been made publicly available, being the first multi-class peach detection dataset. The improved multi-class peach detector based on YOLOv5s by fusing multi-modal images as input and introducing coordinate attention mechanism and depthwise separable convolution was presented.

The experimental comparison results showed that the improved YOLOv5s using multi-modal visual data offered the detection mAP of 98.6% and 88.9% on the naked and bagging peach in complex illumination and severe occlusion environment, increasing by 5.3% and 16.5% than using RGB images, as well as by 2.8% and 6.2% when compared to YOLOv5s. While compared with other networks in detecting bagging peaches, the improved YOLOv5s was the best in terms of mAP, which was 16.3%, 8.1% and 4.5% higher than YOLOX-Nano, PP-YOLO-Tiny, and EfficientDet-D0, respectively. The improved YOLOv5s with multi-modal visual data could enhance the model's perception ability in detecting both naked and bagging peaches under severe occlusion scenes, as well as under various illumination conditions.

In particular, the depth imaging modality could reduce the false and missing detection of peach targets under artificial illumination condition, and the infrared imaging modality could improve the detection accuracy under strong illumination condition.

Additionally, it was found out that the proposed detection model could reach 19 times detection per second with the considered five-channel multi-modal images on popular embedded platform, which could meet the real-time requirement of fruit harvesting system.

The main limitation of using five-channel multi-modal images was the underutilization of spatial geometric information in the depth and infrared images. Future work includes the exploration of stronger fruit detection networks and multi-modal image fusion methods for further improving the detection of the in-field bagging fruits, as well as ones with various types of bags wrapped.

References:

[1] YADAV S, SENGAR N, SINGH A, et al. Identification of disease using deep learning and evaluation of bacteriosis in peach leaf[J]. Ecological Informatics, 2021, 61: ID 101247.

[2] GENE-MOLA J, VILAPLANA V, ROSELL-POLO J R, et al. Multi-modal deep learning for Fuji apple detection using RGB-D cameras and their radiometric capabilities[J]. Computers and Electronics in Agriculture, 2019, 162: 689-698.

[3] NGUYEN T T, VANDEVOORDE K, WOUTERS N, et al. Detection of red and bicoloured apples on tree with an RGB-D camera[J]. Biosystems Engineering, 2016, 146: 33-44.

[4] LIU X, JIA W, RUAN C, et al. The recognition of apple fruits in plastic bags based on block classification[J]. Precision Agriculture, 2018, 19(4): 735-749.

[5] LIU T, EHSANI R, TOUDESHKI A, et al. Identifying immature and mature pomelo fruits in trees by elliptical model fitting in the Cr-Cb color space[J]. Precision Agriculture, 2019, 20(1): 138-156.

[6] LIU Y, CHEN B, QIAO J. Development of a machine vision algorithm for recognition of peach fruit in a natural scene[J]. Transactions of the ASABE, 2011, 54(2): 695-702.

[7] WILLIAMS H A M, JONES M H, NEJATI M, et al. Robotic kiwifruit harvesting using machine vision, convolutional neural networks, and robotic arms[J]. Biosystems Engineering, 2019, 181: 140-156.

[8] NAVAS E, FERNANDEZ R, SEPULVEDA D, et al. Soft grippers for automatic crop harvesting: A review[J]. Sensors, 2021, 21(8): ID 2689.

[9] TU S, PANG J, LIU H, et al. Passion fruit detection and counting based on multiple scale faster R-CNN using RGB-D images[J]. Precision Agriculture, 2020, 21 (5): 1072-1091.

[10] HÄNI N, ROY P, ISLER V. A comparative study of fruit detection and counting methods for yield mapping in apple orchards[J]. Journal of Field Robotics, 2020, 37(2): 263-282.

[11] LU S, CHEN W, ZHANG X, et al. Canopy-attention-YOLOv4-based immature/mature apple fruit detection on dense-foliage tree architectures for early crop load estimation[J]. Computers and Electronics in Agriculture, 2022, 193: ID 106696.

[12] LI X, PAN J, XIE F, et al. Fast and accurate green pepper detection in complex backgrounds via an improved YOLOv4-tiny model[J]. Computers and Electronics in Agriculture, 2021, 191: ID 106503.

[13] JIANG M, SONG L, WANG Y, et al. Fusion of the YOLOv4 network model and visual attention mechanism to detect low-quality young apples in a complex environment[J]. Precision Agriculture, 2022, 23(2): 559-577.

[14] HUANG H, HUANG T, LI Z, et al. Design of citrus fruit detection system based on mobile platform and edge computer device[J]. Sensors, 2021, 22(1): ID 59.

[15] FU L, GAO F, WU J, et al. Application of consumer RGB-D cameras for fruit detection and localization in field: A critical review[J]. Computers and Electronics in Agriculture, 2020, 177: ID 105687.

[16] SA I, GE Z, DAYOUB F, et al. Deepfruits: A fruit detection system using deep neural networks[J]. Sensors, 2016, 16(8): ID 1222.

[17] ARAD B, BALENDONCK J, BARTH R, et al. Development of a sweet pepper harvesting robot[J]. Journal of Field Robotics, 2020, 37(6): 1027-1039.

[18] SUO R, GAO F, ZHOU Z, et al. Improved multi-classes kiwifruit detection in orchard to avoid collisions during robotic picking[J]. Computers and Electronics in Agriculture, 2021, 182: ID 106052.

[19] TIAN Y, YANG G, WANG Z, et al. Apple detection during different growth stages in orchards using the improved YOLO-v3 model[J]. Computers and Electronics in Agriculture, 2019, 157: 417-426.

[20] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, New York, USA: IEEE, 2016: 779-788.

[21] REDMON J, FARHADI A. YOLOv3: An incremental improvement[J/OL]. arXiv:1804.02767[cs.CV], 2018.

[22] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, New York, USA: IEEE, 2017: 7263-7271.

[23] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection[J/OL]. arXiv: 2004.10934[cs.CV], 2020.

[24] YAN B, FAN P, LEI X, et al. A real-time apple targets detection method for picking robot based on improved YOLOv5[J]. Remote Sensing, 2021, 13(9): ID 1619.

[25] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, New York, USA: IEEE, 2018: 8759-8768.

[26] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, New York, USA: IEEE, 2021: 13713-13722.

[27] FANG L, WU Y, LI Y, et al. Ginger seeding detection and shoot orientation discrimination using an improved YOLOv4-LITE network[J]. Agronomy, 2021, 11(11): ID 2328.

[28] SHI C, LIN L, SUN J, et al. A lightweight YOLOv5 transmission line defect detection method based on coordinate attention[C]// 2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC). Piscataway, New York, USA: IEEE, 2022, 6:

1779-1785.
[29] ZHA M, QIAN W, YI W, et al. A lightweight YOLOv4-based forestry pest detection method using coordinate attention and feature fusion[J]. Entropy, 2021, 23(12): 1587.
[30] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, New York, USA: IEEE, 2018: 7132-7141.
[31] ZHANG Y, YU J, CHEN Y, et al. Real-time strawberry detection using deep neural networks on embedded system (RTSD-net): An edge AI application[J]. Computers and Electronics in Agriculture, 2022, 192: ID 106586.
[32] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, New York, USA: IEEE, 2017: 1251-1258.
[33] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection[J/OL]. arXiv: 2004.10934[cs.CV], 2020.
[34] ZHENG Z, WANG P, LIU W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Piscataway, New York, USA: IEEE, 2020, 34(7): 12993-13000.
[35] POWERS D M W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation[J/OL]. arXiv: 2010.16061[cs.LG], 2020.
[36] GE Z, LIU S, WANG F, et al. YOLOx: Exceeding yolo series in 2021[J/OL]. arXiv: 2107.08430[cs.CV], 2021.
[37] LONG X, DENG K, WANG G, et al. PP-YOLO: An effective and efficient implementation of object detector[J/OL]. arXiv: 2007.12099[cs.CV], 2020.
[38] TAN M, PANG R, LE Q V. EfficientDet: Scalable and efficient object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, New York, USA: IEEE, 2020: 10781-10790.

基于改进YOLOv5s和多模态图像的树上毛桃检测

罗　庆[1,2,3]，饶　元[1,2,3*]，金　秀[1,2,3]，江朝晖[1,2,3]，王　坦[1,2,3]，
王丰仪[1,2,3]，张　武[1,2,3]

（1. 安徽农业大学 信息与计算机学院，安徽合肥 230036；2. 农业农村部农业传感器重点实验室，安徽合肥 230036；3. 智慧农业技术与装备安徽省重点实验室，安徽合肥 230036）

摘　要：毛桃等果实的准确检测是实现机械化、智能化农艺管理的必要前提。然而，由于光照不均和严重遮挡，在果园中实现毛桃，尤其是套袋毛桃的检测一直面临着挑战。本研究基于改进YOLOv5s和多模态视觉数据提出了面向机械化采摘的毛桃多分类准确检测。具体地，构建了一个多类标签的裸桃和套袋毛桃的RGB-D数据集，包括4127组由消费级RGB-D相机获取的像素对齐的彩色、深度和红外图像。随后，通过引入方向感知和位置敏感的注意力机制，提出了改进的轻量级YOLOv5s（小深度）模型，该模型可以沿一个空间方向捕捉长距离依赖，并沿另一个空间方向保留准确的位置信息，提高毛桃检测精度。同时，通过将卷积操作分解为深度方向的卷积与宽度、高度方向的卷积，使用深度可分离卷积在保持模型检测准确性的同时减少模型的计算量、训练和推理时间。实验结果表明，使用多模态视觉数据的改进YOLOv5s模型在复杂光照和严重遮挡环境下，对裸桃和套袋毛桃的平均精度（Mean Average Precision，mAP）分别为98.6%和88.9%，比仅使用RGB图像提高了5.3%和16.5%，比YOLOv5s提高了2.8%和6.2%。在套袋毛桃检测方面，改进YOLOv5s的mAP比YOLOX-Nano、PP-YOLO-Tiny和EfficientDet-D0分别提升了16.3%、8.1%和4.5%。此外，多模态图像、改进YOLOv5s对提升自然果园中的裸桃和套袋毛桃的准确检测均有贡献，所提出的改进YOLOv5s模型在检测公开数据集中的富士苹果和猕猴桃时，也获得了优于传统方法的结果，验证了所提出的模型具有良好的泛化能力。最后，在主流移动式硬件平台上，改进后的YOLOv5s模型使用五通道多模态图像时检测速度可达每秒19幅，能够实现毛桃的实时检测。上述结果证明了改进的YOLOv5s网络和含多类标签的多模态视觉数据在实现果实自动采摘系统视觉智能方面的应用潜力。

关键词：多类检测；YOLOv5s；多模态视觉数据；机械化采摘；深度学习